THE AMINO ACID COMPOSITION OF PROTEINS: A METHOD OF ANALYSIS

by

R. Dennis Cook and R. F. Nassar
University of Minnesota, St. Paul
Kansas State University, Manhattan

Technical Report No. 221

November 1973

University of Minnesota
Minneapolis, Minnesota

1. Introduction

Amino acid composition of proteins has recently received much consideration in relation to the Darwinian and non-Darwinian theories of evolution (see for example Kimura and Ohta, 1972; King and Jukes, 1969; Ohta and Kimura, 1971). Much attention has been directed toward predicting the average amino acid composition of proteins from the genetic code by assuming a random arrangement of nucleotide bases within a gene. The results from these studies generally provide some support for the non-Darwinian evolutionary theory. The implication is that many amino acid substitutions in evolution were the result of random fixation of selectively neutral mutants. Although the non-Darwinian model has found support, there are strong indications of non-randomness in the evolution of the amino acid composition of proteins (King and Jukes, 1969; Ohta and Kimura, 1971). The apparent deficiency arginine is one example.

The basic technique that has been used to test the non-Darwinian hypothesis is simply a comparison of the observed and expected amino acid compositions. Unfortunately, because of the ad hoc nature of the methods used to determine expected amino acid composition, it is not at all clear what the complete underlying probability model for these expectations is. Clearly, the method used to obtain expected amino acid composition can influence the conclusions. In addition, very little attempt has been made to estimate the base frequencies in the third codon position. The need for better methods of estimation is of major importance and has been recognized by others (Kimura and Ohta, 1972). Inherently, more refined methods of estimation

cannot be developed without the simultaneous consideration of probability models subject to which the estimation will take place. Recent investigations have been concerned with 'independence' between the codon positions. Many other possible models are clearly available.

It has also been argued that present information leads more simply to a neo-Darwinian model and that observed amino acid frequencies and, thus, the genetic code are the result of the effects of selection (Richmond, 1970). Problems of this nature cannot begin to be resolved without additional information, and the information available from amino acid data cannot be effectively understood without the aid of well-defined probability models and refined methods of analysis.

In this paper we discuss a class of loglinear models for describing the relationships between the amino acid composition of proteins and the structure of the genetic code. The associated maximum likelihood procedure that allows for estimation of base frequencies in all three codon positions is also discussed. The technique proposed by Ohta and Kimura (1970) for estimating base frequencies is investigated from the maximum likelihood point of view and , finally, an example is presented.

2. Structure of the RNA Code Table

For the purposes of estimating the base composition of RNA which corresponds to a protein molecule, the standard RNA code table (see Table I) can be visualized as a 4x4x4 contingency table: each of the 64 cells in the table corresponding to a particular codon in the RNA molecule. Given observed amino acid frequencies the problem is to estimate the frequency of each of the 64 codons assuming some model (e.g. 'independence') for the individual cell probabilities. Because, generally, multiple codons code for the same amino acid there is considerable indeterminancy in the code table. For example, the frequencies of the six codons which code for serine will be mixed-up since only the total frequency of an amino acid can be determined. The frequencies of the individual codons are, for the most part, indeterminate. The problem is further confounded by the three chain terminating codons which represent *a priori* zeros in the table. Such zeros are referred to as structural zeros to distinguish them from cells containing observed zeros as a result of sampling variation (Fienberg, 1972). A contingency table containing structural zeros is said to be incomplete.

The problem may now be considered as one of estimating individual cell probabilities in a 4x4x4 incomplete contingency table with mixed-up frequencies. A general maximum likelihood (ML) solution to this type of problem has been obtained by T. Chen (1972) and the method discussed here is basically an adaptation of this method.

From inspection of Table I it is seen that for all possible compositions of the first and second codon positions it is impossible to distinguish

between U and C in the third codon position. For example, in attempting to classify the amino acid His it can be determined with certainty that C (A) must occupy the first (second) codon position. The third codon position may be either U or C. This same type of argument is true for each of the twenty amino acids. Thus, unless some simplifying assumptions are made (e.g., the marginal frequencies of U and C in the third codon position are the same) no estimation procedure will be able to distinguish between U and C in this position. Therefore, we combine U and C in the third codon position to produce Table II. The estimation procedure we consider will be directed toward estimating the individual cell probabilities in Table II; that is, we will restrict attention to estimating the combined frequency of U and C in the third codon position. Once estimates of the combined frequencies are obtained they can be allocated to U and C in any manner desireable.

Let

P_{ijk} = frequency of the codon with base i (i = 1, 2, 3, 4) in the first codon position, base j (j = 1, 2, 3, 4) in the second codon position, and base(s) k (k = 1, 2, 3,) in the third codon position.

In the first and second codon positions the subscript values (i.e. i = 1, 2, 3, 4) correspond to U, C, A, and G, respectively. In the third codon position the subscript values correspond to U + C, A, and G, respectively. For example,

$$P_{123} = \Pr(\underline{U}\ \underline{C}\ \underline{G})$$

$$P_{341} = \Pr(\underline{A}\ \underline{G}\ \underline{U}) + \Pr(\underline{A}\ \underline{G}\ \underline{C})$$

Because of the three chain-terminating codons any estimation procedure must

preserve the property that $P_{132} = P_{133} = P_{142} = 0$. In subsequent sections it will be necessary to distinguish between the set (S) of cells not containing structural zeros and the set of three cells containing structural zeros.

Let

$$P_{ij+} = \sum_{k=1}^{3} P_{ijk}$$

$$P_{i++} = \sum_{j} \sum_{k} P_{ijk}$$

with similar definitions holding for P_{+j+}, P_{++k}, P_{i+k}, and P_{+jk}. Notice that P_{i++}, P_{+j+}, and P_{++k} denote the marginal frequency distributions of the first, second, and third codon positions, respectively.

3. Likelihood

Here we consider the likelihood function from which the maximum likelihood estimates will be obtained.

Let

n_i = The observed frequency of the ith amino acid, $i = 1, 2, \ldots 20$

$N = \sum n_i$

f_i = The probability of observing the ith amino acid, $i = 1, 2, \ldots 20$.

The probabilities, f_i, are linear functions of the individual cell probabilities, P_{ijk}, as given in Table III. Notice also that Table III relates each n_i to a particular amino acid. The amino acids were indexed in a convenient manner by beginning in the upper left-hand corner of Table II and labeling down the

first column, etc.

The probability distribution function of the observations, $\{n_i\}$, is multinomial:

$$f(n_1,\ldots,n_{20}) = \frac{N!}{\prod_{i=1}^{20} n_i} \prod_{i=1}^{20} [f_i]^{n_i} \qquad (1)$$

subject to the condition $\Sigma f_i = \Sigma P_{ijk} = 1$.

Using (1) it will be impossible to estimate each P_{ijk} subject only to the constraint $\Sigma P_{ijk} = 1$, since there are at most 19 free parameters or degrees of freedom (corresponding to the twenty amino acids minus one for the additive constraint on the cell probabilities) and estimating each cell probability would require many more. Thus, we next consider a class of models for the individual cell probabilities on which the estimation will be based.

4. Models

Here we consider the class of hierarchical loglinear models (Fienberg, 1970) for the underlying cell probabilities. Strictly speaking, the class of models to be considered here is termed quasi-loglinear to distinguish them from the models used for contingency tables not containing structural zeros. Under these models each cell probability is expressed as

$$\log P_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(kj)} + u_{123(ijk)} \tag{2}$$

for $(ijk) \in S$ and for this problem subject to the constraints

$$\sum_i u_{1(i)} = \sum_i u_{12(ij)} = \sum_i u_{13(ik)} = \sum_i \delta_{ijk}\, u_{123(ijk)} = 0 \tag{3a}$$

$$\sum_j u_{2(j)} = \sum_j u_{12(ij)} = \sum_j u_{23(jk)} = \sum_j \delta_{ijk}\, u_{123(ijk)} = 0 \tag{3b}$$

$$\sum_k u_{3(k)} = \sum_k u_{13(ik)} = \sum_k u_{23(jk)} = \sum_k \delta_{ijk}\, u_{123(ijk)} = 0 \tag{3c}$$

where $\delta_{ijk} = \begin{cases} 1 & \text{if } (ijk) \in S \\ 0 & \text{elsewhere} \end{cases}$

These are a special case of the general constraints for quasi-loglinear models (Fienberg, 1972) and arise because there are no structural zeros in the two-dimensional marginal tables obtained by collapsing Table II. Taking the above constraints into account, it is easily seen that the number of free parameters required for the terms u, u_1, u_2, u_3, u_{12}, u_{13}, and u_{23} are 1, 3, 3, 2, 3x3, 3x2, and 3x2, respectively. The number of free parameters in u_{123} is 3x3x2 minus the number of structural zeros, or 15. Thus the full model (2) requires the estimation of 45 free parameters.

The class of models to be considered here is the one obtained by the setting of u-terms in (2) to zero. As mentioned previously, the full model (2) cannot be considered because of the restricted number of degrees of freedom available. The model of no second-order interaction among the individual cell probabilities is obtained by setting $u_{123(ijk)} = 0$ for $(ijk) \in S$. This model requires 45 - 15 = 30 free parameters and this is again too many to be accommodated by the 19 free parameters available. Continuing to set u-terms to zero in a hierarchical fashion we find that we can have at most one first-order interaction term in the model. Thus we can only consider the following four models:

Model 1

$$\log P_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}, \quad (ijk) \in S \tag{4}$$

Model 2

$$\log P_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)}, \quad (ijk) \in S \tag{5}$$

Model 3

$$\log P_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)}, \quad (ijk) \in S \tag{6}$$

Model 4

$$\log P_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}, \quad (ijk) \in S \tag{7}$$

Model 1 is obtained by setting all first and second-order interaction terms in (2) to zero. The individual u-terms in (4) are subject to the applicable constraints given in (3a), (3b), and (3c); namely, $\sum_i u_{1(i)} = \sum_j u_{2(j)} = \sum_k u_{3(k)} = 0$. Thus, this model requies that only nine free parameters be fitted: one for u, three for each of $u_{1(i)}$ and $u_{2(j)}$, and two for $u_{3(k)}$.

From (4) it is easily seen that model 1 is just a convenient way of writing

$$P_{ijk} = a_i b_j c_k \qquad (ijk) \in S \tag{8}$$

for positive constants a_i, b_j and c_k. If there were no structural zeros in the code table this model would be the same as the familiar model of independence and a little algebra would verify that the individual cell probabilities satisfy

$$P_{ijk} = P_{i++} P_{+j+} P_{++k} \tag{9}$$

for all i, j and k. However, because of the chain terminating codons the three codon positions cannot be mutually independent (i.e., $P_{1++}P_{+3+}P_{++2} \neq P_{132} = 0$) and relation (9) is no longer valid. To distinguish between models of independence characterized by (8), and models of independence usually associated with contingency tables (i.e., those satisfying (9)), the former is referred to as a model of 'quasi-independence' (Fienberg, 1972). Quasi-independence is a form of independence conditional on the restriction of our attention to that portion of Table II not containing the three chain-terminating codons. Each of the four models given above can be interpreted

as a model of quasi-independence in one way or another.

Model 2 contains a single first-order interaction term and is also subject to the applicable conditions in (3a), (3b) and (3c). In this case the model requires eighteen free parameters; nine as in model 1 plus nine for the u_{12} terms. This model implies that the individual cell probabilities are the product of a parameter depending on the first and second codon positions combined, and another depending on the third codon position:

$$P_{ijk} = a_{ij}\, b_k \qquad (ijk) \in S \tag{10}$$

If the individual cell probabilities were all positive then it could be easily verified that

$$P_{ijk} = P_{ij+}\, P_{++k} \tag{11}$$

for all i, j, and k. Thus, this is a model of quasi-independence between the first and second codon positions together, and the third codon position. Models 3 and 4 have interpretations analogous to that of model 2 and each required fifteen free parameters.

It must be mentioned that any one of the above four models may be insufficient to completely describe a set of amino acid data. However, once the class of quasi-loglinear models is adopted our attention must be confined to these models because of the restricted number of degrees of freedom available.

We now consider the ML estimates of individual cell probabilities based on the above models.

5. ML Estimates

Under the multinomial distribution assumption and for any of the four models the log-likelihood function is obtained from (1). ML estimates, $\hat{P}_{ijk}$, of individual cell probabilities can be found by substituting for each f_i in the log-likelihood individual cell probabilities from Table III and transforming each P_{ijk} using either (4), (5), (6), or (7), whichever is appropriate. The resulting transformed log-likelihood can be differentiated with respect to each u-parameter in the model and the resulting set of ML equations solved simultaneously for the ML estimates. This procedure is rather combersome. Fortunately, following Chen (1972), the ML equations are easily written down once the general pattern is recognized. Symbolically, the set of ML equations for each of the four models in question can be written as follows:

Model 1

$$E_{\hat{p}}[m_1(i)|\underline{n}] = N \sum_j \sum_k \delta_{ijk} \hat{P}_{ijk}, \qquad i = 1, 2, 3, 4 \tag{12}$$

$$E_{\hat{p}}[m_2(j)|\underline{n}] = N \sum_i \sum_k \delta_{ijk} \hat{P}_{ijk}, \qquad j = 1, 2, 3, 4 \tag{13}$$

$$E_{\hat{p}}[m_3(k)|\underline{n}] = N \sum_i \sum_j \delta_{ijk} \hat{P}_{ijk}, \qquad k = 1, 2, 3 \tag{14}$$

Model 2 - Equation (14) and

$$E_{\hat{p}}[m_{12}(ij)|\underline{n}] = N \sum_k \delta_{ijk} \hat{P}_{ijk} \qquad \begin{matrix} i = 1, 2, 3, 4 \\ j = 1, 2, 3, 4 \end{matrix} \tag{15}$$

Model 3 - Equation (13) and

$$E_{\hat{p}}[m_{13}(ik)|\underline{n}] = N \sum_{j} \delta_{ijk} \hat{P}_{ijk} \qquad \begin{matrix} i = 1, 2, 3, 4 \\ k = 1, 2, 3 \end{matrix} \qquad (16)$$

Model 4 - Equation (12) and

$$E_{\hat{p}}[m_{23}(jk)|\underline{n}] = N \sum_{i} \delta_{ijk} \hat{P}_{ijk} \qquad \begin{matrix} j = 1, 2, 3, 4 \\ k = 1, 2, 3. \end{matrix} \qquad (17)$$

In the above equations, $\underline{n}$ represents the vector of observed amino acid frequencies; m_1, m_2, m_3, m_{12}, m_{13}, and m_{23} denote the marginal counts for the indicated codon position(s); and $E_{\hat{p}}$ denotes the expectation of marginal counts given the observed amino acid frequencies.

The left hand side of each of the above equations represents the sum of actual observations which fall in the margin in question plus the proportion of the mixed-up frequencies expected to fall in the marginal position. A few examples will make this clear. Consider first the marginal expectation for $\underline{U}$ in the first codon position,

$$E_{\hat{p}}[m_1(1)|\underline{n}] = \sum_{j} \sum_{k} \delta_{ijk} \hat{P}_{ijk}$$

to determine the LHS of this equation first note that the actual nonmixed-up counts for uracil in the first codon position are n_1, n_{10}, n_{17}, and n_{18}. These counts correspond to the four amino acids Phe, Tyr, Cys, and Try respectively. Second, an observation will be Leu with probability $f_2 = P_{112} + P_{113} + P_{211} + P_{212} + P_{213}$ and Ser with probability f_6. The proportion of observation classified as Leu that is expected to fall in cells (1, 1, 2) and (1, 1, 3) is simply $n_2(P_{112} + P_{113})/f_2$ and the proportion of those observations classified as Ser expected to fall in cells

(1, 2, 1), (1, 2, 2) and (1, 2, 3) is simply $n_6(P_{121} + P_{122} + P_{123})/f_6$. Thus, we have

$$E_{\hat{p}}[m_1(1)|\underline{n}] = n_1 + n_{10} + n_{17} + n_{18} + n_2(\hat{P}_{112} + \hat{P}_{113})/\hat{f}_2$$

$$+ n_6(\hat{P}_{121} + \hat{P}_{122} + \hat{P}_{123})/\hat{f}_6$$

$$= \sum_j \sum_k \delta_{ijk} \hat{P}_{ijk} .$$

Similarly, we have

$$E_{\hat{p}}[m_2(1)|\underline{n}] = n_1 + n_2 + n_3 + n_4 + n_5$$

$$E_{\hat{p}}[m_2(4)|\underline{n}] = n_6 \hat{P}_{341}/\hat{f}_6 + n_{17} + n_{18} + n_{19} + n_{20} .$$

The following three examples should be an aid to computing expectations for the two dimensional margins.

$$E_{\hat{p}}[m_{12}(1, 1)|\underline{n}] = n_1 + n_2(\hat{P}_{112} + \hat{P}_{113})/\hat{f}_2$$

$$E_{\hat{p}}[m_{12}(3, 1)|\underline{n}] = n_3 + n_4$$

$$E_{\hat{p}}[m_{13}(1, 1)|\underline{n}] = n_6 \hat{P}_{121}/\hat{f}_6 + n_1 + n_{10} + n_{17} .$$

It appears that no closed-term solution to the ML equations exists for any of the four models above and the actual solution of each set of ML equations must be determined by iterative methods. The following two iterative methods are extensions of the Deming-Stephen iterative proportional fitting procedure

(Chen, 1972; Fienberg, 1972; Goodman, 1968) and can be used to determine the solution of the ML equations. Method 1 is applicable to model 1 and method 2 is applicable to models 2, 3, and 4.

Method 1 - Let $P_{ijk}^{(r)}$ denote the estimates of the individual cell probabilities produced on the rth iteration. Choose $P_{ijk}^{(0)} = \delta_{ijk}$ and define

$$S_i^{(r)} = \sum_j \sum_k \delta_{ijk} P_{ijk}^{(r)} \qquad i = 1, 2, 3, 4$$

$$S_j^{(r)} = \sum_i \sum_k \delta_{ijk} P_{ijk}^{(r)} \qquad j = 1, 2, 3, 4$$

$$S_k^{(r)} = \sum_i \sum_j \delta_{ijk} P_{ijk}^{(r)} \qquad k = 1, 2, 3 .$$

The (r+1)st iteration for this method consists of the following six steps:

$$S_i^{(r+1)} = \frac{1}{N} E_{p^{(r)}} [m_1(i) | \underline{n}] \tag{18}$$

$$P_{ijk}^{(r+1)} = S_i^{(r+1)} P_{ijk}^{(r)} / S_i^{(r)} \tag{19}$$

$$S_j^{(r+2)} = \frac{1}{N} E_{p^{(r+1)}} [m_2(j) | \underline{n}] \tag{20}$$

$$P_{ijk}^{(r+2)} = S_j^{(r+2)} P_{ijk}^{(r+1)} / S_j^{(r+1)} \tag{21}$$

$$S_k^{(r+3)} = \frac{1}{N} E_{p^{(r+2)}} [m_3(k) | \underline{n}] \tag{22}$$

$$P_{ijk}^{(r+3)} = S_k^{(r+3)} P_{ijk}^{(r+2)} / S_k^{(r+2)} \tag{23}$$

Iteration is continued until the desired accuracy is achieved.

Method 2 - This method will be illustrated using model 2. The adaptation to models 3 and 4 is straightforward. Let $P_{ijk}^{(r)}$ be defined as for method 1. Define

$$S_{ij}^{(r)} = \sum_k \delta_{ijk} P_{ijk}^{(r)} \qquad i, j = 1, 2, 3, 4$$

The (r+1)st iteration consists of the following four steps:

$$S_{ij}^{(r+1)} = \frac{1}{N} E_{P^{(r)}}[m_{12}(ij) | \underline{n}] \tag{24}$$

$$P_{ijk}^{(r+1)} = S_{ij}^{(r+1)} \; P_{ijk}^{(r)} \; / \; S_{ij}^{(r)} \tag{25}$$

$$S_{k}^{(r+2)} = \frac{1}{N} E_{P^{(r+1)}}[m_{3}(k) | \underline{n}] \tag{26}$$

$$P_{ijk}^{(r+2)} = S_{k}^{(r+2)} \; P_{ijk}^{(r+1)} \; / \; S_{k}^{(r+1)} \tag{27}$$

Again, iteration is continued until the desired accuracy is achieved.

After the ML estimates of the individual cell probabilities, $\hat{P}_{ijk}$, are obtained, the ML estimates of amino acid frequencies, $\hat{f}_k$, can be found by adding the appropriate individual cell probabilities (see Table III). The goodness-of-fit for each model can then be tested using the Pearson Chi-square statistic

$$\chi^2 = \sum_{i=1}^{20} \frac{(n_i - N\hat{f}_i)^2}{N\hat{f}_i} \tag{28}$$

where N is the total number of amino acids counted, or the corresponding likelihood ratio test statistic

$$L^2 = 2 \sum_{i=1}^{20} n_i \log(n_i / N\hat{f}_i) \tag{29}$$

Under the null hypothesis (i.e. the model in question) both χ^2 and L^2 have asymptotic central chi-square distributions with degrees of freedom equal to twenty minus the number of free parameters fitted for the model in question, i.e. the degrees of freedom are 11, 2, 5, and 5 for models 1, 2, 3, and 4, respectively.

The presence of structural zeros in the code table makes interpretation of these models more difficult than those usually associated with contingency tables. Unless considerable care is exercised, erroneous conclusions will result. The remaining discussion serves to point out the major pitfalls to be avoided.

After deciding that one of the four possible models adequately describes a set of amino acid data, a natural tendency is to collapse the code table and investigate selected one- and two-dimensional tables. For example, if model 2 is chosen, the marginal table for the first and second codon positions obtained by adding over the third codon position might be used to investigate the nature of the interaction. A point to remember is that because of the structural zeros, relationships that hold in the three-dimensional table may not hold in the marginal tables. As a simple illustration suppose that model 1 perfectly describes a set of data. The marginal table for the first and second positions contains no structural zeros and, thus, it might be thought that these codon positions would be independent in the manner usually associated with contingency tables. That this is not true can be seen as follows: Quasi-independence implies that

$$P_{ijk} = a_i b_j c_k \qquad (ijk) \in S$$

where, without loss of generality, we can take $\Sigma\, c_k = 1$. The marginal distribution for the first and second positions is

$$P_{ij+} = a_i b_j \qquad (ij) \neq (13) \text{ or } (14)$$
$$P_{13+} = a_1 b_3 c_1$$
$$P_{14+} = a_1 b_4 (c_1 + c_3)$$

The relationship required for independence does not hold and, thus, the first and second codon positions are not independent.

Similar reasoning can be used as an aid to understanding the structure of marginal tables when the model for the full table contains an interaction term. Generally, relationships that are found to hold in the full code table cannot be assumed true for the marginal tables and vice versa.

6. The Method of Ohta and Kimura

Ohta and Kimura (1970) described a method of estimating the base frequencies in the first and second codon positions under the model of 'independence' between these positions (see also Kimura and Ohta, 1972). Because of the emphasis that has been placed on the results obtained using this method, we feel it is important to understand exactly what they propose. In this section we offer a derivation and discussion of this method.

Since the method of Ohta and Kimura was presented to estimate base frequencies in the first and second codon positions, we consider only the joint marginal distribution for these positions (see Table IV). Notice that in Table IV the observations in the set of cells, $S_1 = \{(1,1),(2,1)\}$, are partially mixed-up, and the observations in the set of cells, $S_2 = \{(1,2),(2,4),(3,4)\}$, are partially mixed-up. Let $S = S_1 + S_2$, and define

m_{ij} = observed number of nonmixed-up observations falling in cells $(i, j) \notin S$

P_{ij} = joint probability of having base i in codon position 1 and base j in codon position 2.

$P_{i+} = \sum_{j} P_{ij}$

$P_{+j} = \sum_{i} P_{ij}$

Notice that each m_{ij} will be either the observed frequency of an amino acid or the sum of the frequencies for two amino acids. Let n_i $(i = 1, 2...20)$ be defined as in Table III. We now proceed to establish the likelihood function of the observations $(m_{ij}$ $(ij) \notin S$, n_1, n_2, n_6, $n_{19})$ and the estimates of the individual cell probabilities under the model of independence.

That portion of the likelihood corresponding to the nonmixed-up cells is proportional to

$$\prod_{(ij)\notin S} (P_{ij})^{m_{ij}} \tag{30}$$

Under the hypothesis of independence each individual cell probability is the product of the appropriate marginal frequencies $(P_{ij} = P_{i+} P_{+j})$ and thus, (30) becomes

$$\prod_{(ij)\notin S} (P_{i+} P_{+j})^{m_{ij}} \tag{31}$$

To establish that part of the likelihood corresponding to S, first, define

λ_1 = Probability an observation with base composition UU in the first two codon positions is Phe.

λ_2 = Probability an observation with base composition AG in the first two codon positions is Ser.

It follows immediately from the above definitions that $1 - \lambda_1$ is the probability on observation with base composition UU is Leu and $1 - \lambda_2$ is the probability on observation with base composition AG is Arg. Now, the parts of the likelihood corresponding to the n_1 Phe observations and n_2 Leu observations are

$$(P_{1+}\ P_{+1}\ \lambda_1)^{n_1} \tag{32}$$

and

$$[P_{1+}\ P_{+1}(1 - \lambda_1) + P_{2+}\ P_{1+}]^{n_2} \tag{33}$$

Analogously, the parts of the likelihood corresponding to the n_6 Ser observations and n_{19} Arg observations are

$$[P_{1+}\ P_{+2} + P_{3+}\ P_{+4}\ \lambda_2]^{n_6} \tag{34}$$

and

$$[P_{2+}\ P_{4+} + P_{3+}\ P_{+4}(1 - \lambda_2)]^{n_{19}} \tag{35}$$

The total likelihood is, of course, multinomial and is given by the product of (31), (32), (33), (34), and (35).

The ML estimates of the marginal probabilities ($\hat{P}_{i+}$, $\hat{P}_{+j}$) are obtained in the obvious manner by differentiating the log-likelihood function with respect to each parameter and simultaneously solving the resulting ML

equations. The important point here is that the ML estimates will be functions of λ_1 and λ_2. The method proposed by Ohta and Kimura is a special case of the one outlined here and is obtained by a priori choosing $\lambda_1 = \lambda_2 = \frac{1}{2}$. This can be easily verified by carrying out the indicated algebra.

Under the model of independence λ_1 and λ_2 can be estimated jointly with the marginal probabilities by differentiating the log-likelihood with respect to λ_1 and λ_2 and obtaining two additional ML equations.

The ML estimates, with λ_1 and λ_2 unrestricted, of the marginal probabilities can be shown, after some simplification, to be the solution to the following set of equations

$$n_1 + n_2 + m_{31} + m_{41} = N \hat{P}_{+1} \tag{36}$$

$$\frac{(n_6 + n_{19})\ \hat{P}_{1+}\ \hat{P}_{+2}}{D} + m_{22} + m_{32} + m_{42} = N \hat{P}_{+2} \tag{37}$$

$$m_{13} + m_{23} + m_{33} + m_{43} = N \hat{P}_{+3} \tag{38}$$

$$\frac{(n_6 + n_{19})(\hat{P}_{2+}\ \hat{P}_{+4} + \hat{P}_{3+}\ \hat{P}_{+4})}{D} + m_{14} + m_{44} = N \hat{P}_{+4} \tag{39}$$

$$\frac{(n_6 + n_{19})\ \hat{P}_{1+}\ \hat{P}_{+2}}{D} + \frac{(n_1 + n_2)\ \hat{P}_{1+}}{\hat{P}_{1+} + \hat{P}_{2+}} + m_{13} + m_{14} = N \hat{P}_{1+} \tag{40}$$

$$\frac{(n_6 + n_{19})\ \hat{P}_{2+}\ \hat{P}_{+4}}{D} + \frac{(n_1 + n_2)\ \hat{P}_{2+}}{\hat{P}_{1+} + \hat{P}_{2+}} + m_{22} + m_{23} = N \hat{P}_{2+} \tag{41}$$

$$\frac{(n_6 + n_{19})\ \hat{P}_{3+}\ \hat{P}_{+4}}{D} + m_{31} + m_{32} + m_{33} = N \hat{P}_{3+} \tag{42}$$

$$m_{41} + m_{42} + m_{43} + m_{44} = N \hat{P}_{4+} \tag{43}$$

where $D = \hat{P}_{1+} \hat{P}_{+2} + \hat{P}_{2+} \hat{P}_{+4} + \hat{P}_{3+} \hat{P}_{+4}$. Any three of the equations (36) - (39) plus any three of the equations (40) - (43) may be used to obtain the marginal probability estimates because of the restriction, $\Sigma P_{i+} = \Sigma P_{+j} = 1$.

Given the estimates of the marginal probabilities, the ML estimates of λ_1 and λ_2 are given by

$$\hat{\lambda}_1 = \frac{n_1(\hat{P}_{1+} + \hat{P}_{2+})}{\hat{P}_{1+}(n_1 + n_2)} \tag{44}$$

$$\hat{\lambda}_2 = \frac{\hat{P}_{1+} \hat{P}_{+2}[n_6 + n_{19}]}{\hat{P}_{1+} \hat{P}_{2+} + \hat{P}_{2+} \hat{P}_{+4} + \hat{P}_{3+} \hat{P}_{+4}} \tag{45}$$

It is important to note that $\hat{\lambda}_1$ and $\hat{\lambda}_2$ were obtained assuming the model of independence. If the first and second codon positions are not, in fact, independent than these estimates will no longer be appropriate. However, in this event estimates of λ_1 and λ_2 can be easily obtained from the analysis of the three dimensional table discussed previously.

Finally, it must be pointed out that analyzing just the two-dimensional table for the first and second codon positions can lead to misleading results in the manner described at the end of Section 5.

7. Example

In this section we present an analysis of the amino acid composition of human cytochrome c in some detail to demonstrate the applicability of the method. Individual cell probability estimates were obtained for models 1 and 2 using the iterative procedures previously outlined. Iteration was carried out on the University of Minnesota CDC 6400 and terminated the first time $|P_{ijk}^{(r)} - P_{ijk}^{(r+10)}| \leq 10^{-8}$.

Table V presents estimates of the probabilities of occurrence of the RNA codons, in the order presented in Table II, under models 1 and 2. Estimates of the marginal frequencies of the bases (Table VI) and the expected frequencies of each amino acid (Table VII) are obtained from Table V by adding the appropriate individual cell probabilities. It is noted that model 1 does not fit the data, since the chi-square of 31.418 with 11 degrees of freedom is highly significant. On the other hand, the fit to model 2 is quite good as judged by the small chi-square value (Table VII). The indication is then that the assumption of independence between bases in codon positions 1 and 2 is incorrect and that a definite relationship exists between these two adjacent codon positions. The discrepancy from quasi-independence seems to be due mainly to the amino acids arginine, glycine, tyrosine, serine, proline, and valine as judged by the magnitude of their deviations from expectations.

The estimates of bases in each codon position show some variation, but not by much, between models 1 and 2. They show that A is much larger than U in positions 1 and 2 and in all likelihood in position 3 also. This is in agreement with what was reported by Ohta and Kimura (1970) for positions 1 and 2. There is an indication also that G is larger than C in all three

positions. A very interesting difference occurs between models 1 and 2 with regard to the frequencies of the leucine codons. Under model 2, of the six codons that could code for leucine, four (CUU, CUC, CUA, CUG) do not seem to exist as indicated by their near zero frequency estimates (Table V). Actually, these estimates would eventually converge to zero if iteration was carried out long enough.

We can explain how these zero estimates arise by an argument that can generally serve as a check on the iterative procedure. Ignoring the structural zeros and under model 2 we have (the subscript indicates the codon position)

$$\Pr(U_1,U_2,(U+C)_3) = \Pr(U_1,U_2)\ \Pr((U+C)_3)$$

and

$$\Pr(U_1,U_2) = \frac{\Pr(U_1,U_2,(U+C)_3)}{\Pr((U+C)_3)} \quad . \tag{46}$$

Given an estimate $\hat{\Pr}(U_1,U_2)$ of $\Pr(U_1,U_2)$, the number expected to fall in cell U_1U_2 of Table II is $N\ \hat{\Pr}(U_1,U_2)$. If

$$N\ \hat{\Pr}(U_1,U_2) \geq n_1 + n_2$$

then all the leucine observations will be expected to fall in U_1U_2 and none in C_1U_2, hence the zero estimate of the latter.

From Tables II and III

$$\hat{\Pr}(U_1,U_2,(U+C)_3) = n_1/N \tag{47}$$

and three independent estimates of $\Pr((C+U)_3)$ are

$$\frac{n_{11}}{n_{11}+n_{12}} \quad , \quad \frac{n_{13}}{n_{13}+n_{14}} \quad , \quad \text{and} \quad \frac{n_{15}}{n_{15}+n_{16}} \quad .$$

These are combined into a weighted average to give

$$\hat{\Pr}((C+U)_3) = \frac{n_{11}+n_{13}+n_{15}}{\sum_{i=11}^{16} n_i} \quad . \tag{48}$$

From (46), (47), and (48) we have

$$\hat{Pr}(U_1,U_2) = (n_1/N)\frac{\sum_{i=11}^{16} n_i}{n_{11} + n_{13} + n_{15}}$$

and

$$N\,\hat{Pr}(U_1,U_2) = \frac{n_1 \sum_{i=11}^{16} n_i}{n_{11} + n_{13} + n_{15}} \quad .$$

Applying this to the cytochrome c data we obtain

$$N\,\hat{Pr}(U_1,U_2) = (3)(39)/11 = 10.6 > n_1 + n_2 = 9$$

which indicates that the cell C_1U_1 is an empty cell with regard to leucine and hence its probability of occurrence is zero. It has been found empirically that this ad hoc technique produces estimates of $Pr(U_1,U_2)$ that are quite close to the ML procedure. In the event that

$$N\,\hat{Pr}(U_1,U_2) < n_1 + n_2$$

the expected frequency will be equal to the observed frequency for both phenylalanine and leucine.

The method of analysis presented in this paper enables us for the first time to test for correlations or interactions between bases in any two codon positions. It would be of considerable interest now to analyze proteins from a wide variety of animals, plants, bacteria, and viruses. From such analysis relationships among base positions in a codon that are not known to us might be revealed.

TABLE I

Standard RNA code table

Ala = Alanine; Arg = Arginine; Asn = Asparagine; Asp = Aspartic acid; Cys = Cysteine; Gln = Glutamine; Glu = Glutamic acid; Gly = Glycine; His = Histidine; Ile = Isoleucine; Leu = Leucine; Lys = Lysine; Met = Methionine; Phe = Phenylalanine; Pro = Proline; Ser = Serine; Thr = Threonine; Try = Tryptophan; Tyr = Tyrosine; Val = Valine; Term. = Chain terminating codon.

1 \ 2	U	C	A	G	2 / 3
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Term.	Term.	A
	Leu	Ser	Term.	Try	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

TABLE II

RNA code table with uracil and cytosine combined in the third codon position.

1 \ 2	U	C	A	G	3 \ 2
U	Phe	Ser	Tyr	Cys	U+C
	Leu	Ser	Term.	Term.	A
	Leu	Ser	Term.	Try	G
C	Leu	Pro	His	Arg	U+C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U+C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U+C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

TABLE III

Expected amino acid frequencies, f_i, in terms of individual cell probabilities, and observed counts n_i.

Amino Acid	Count	Frequency
Phe	n_1	$f_1 = P_{111}$
Leu	n_2	$f_2 = P_{112}+P_{113}+P_{21+}$
Ile	n_3	$f_3 = P_{311}+P_{312}$
Met	n_4	$f_4 = P_{313}$
Val	n_5	$f_5 = P_{41+}$
Ser	n_6	$f_6 = P_{12+}+P_{341}$
Pro	n_7	$f_7 = P_{22+}$
Thr	n_8	$f_8 = P_{32+}$
Ala	n_9	$f_9 = P_{42+}$
Tyr	n_{10}	$f_{10} = P_{131}$
His	n_{11}	$f_{11} = P_{231}$
Gln	n_{12}	$f_{12} = P_{232}+P_{233}$
Asn	n_{13}	$f_{13} = P_{331}$
Lys	n_{14}	$f_{14} = P_{332}+P_{333}$
Asp	n_{15}	$f_{15} = P_{431}$
Glu	n_{16}	$f_{16} = P_{432}+P_{433}$
Cys	n_{17}	$f_{17} = P_{141}$
Try	n_{18}	$f_{18} = P_{143}$
Arg	n_{19}	$f_{19} = P_{24+}+P_{342}+P_{343}$
Gly	n_{20}	$f_{20} = P_{44+}$

TABLE IV

Marginal RNA code table for the first and second codon positions only.

1 \ 2	U	C	A	G
U	Phe + Leu	Ser	Tyr	Cys + Try
C	Leu	Pro	His + Gln	Arg
A	Ile + Met	Thr	Asn + Lys	Arg + Ser
G	Val	Ala	Asp + Glu	Gly

Table V

Maximum Likelihood Estimates of the Frequencies of the RNA Codons in the Order Presented in Table II for (a) Model 1 and (b) Model 2.

(a) Model 1

0.01390	0.01099	0.02864	0.01473
0.01424	0.01126	0	0
0.02028	0.01604	0	0.02149
0.006736	0.005329	0.01388	0.007139
0.006902	0.005460	0.01422	0.007315
0.009828	0.007775	0.02025	0.01041
0.02430	0.01922	0.05008	0.02575
0.02490	0.01969	0.05131	0.02638
0.03545	0.02804	0.07306	0.03757
0.01854	0.01467	0.03821	0.01965
0.01900	0.01503	0.03915	0.02013
0.02705	0.02140	0.05575	0.02867

(b) Model 2

0.02376	0.004371	0.009615	0.03006
0.03332	0.006129	0	0
0.02944	0.005416	0	0.03724
3.204×10^{-12}	0.01056	0.01320	0.002879
4.492×10^{-12}	0.01481	0.01851	0.004036
3.970×10^{-12}	0.01308	0.01635	0.003567
0.02904	0.01848	0.06073	0.003312
0.04072	0.02591	0.08516	0.004644
0.03599	0.02290	0.07525	0.004103
0.007922	0.01584	0.02904	0.03433
0.01110	0.02221	0.04072	0.04813
0.009815	0.01963	0.03599	0.04253

Table VI

Estimates of Marginal Frequencies Obtained from Table V(a,b) for Each of the Four Bases in Codon Positions 1 and 2 and for A, G, U+C in Codon Position 3

Base Frequency

	1st Position				2nd Position				3rd Position		
Model	U	C	A	G	U	C	A	G	U+C	A	G
1	.1516	.1152	.4158	.3173	.2211	.1749	.3846	.2192	.3118	.2750	.4131
2	.1793	.0970	.4062	.3173	.2211	.1793	.3846	.2148	.2931	.3554	.3513

Table VII

Observed and Expected Frequencies of the 20 Amino Acids Under Models 1 and 2 with their Corresponding Chi-Squares

	Model 1		Model 2	
	Frequency		Frequency	
Amino Acid	Observed	Expected	Observed	Expected
Phenylalanine	3	1.445	3	2.471
Leucine	6	6.031	6	6.528
Isoleucine	8	5.116	8	7.256
Methionine	3	3.687	3	3.743
Valine	3	6.718	3	3.
Serine	2	6.662	2	2.
Proline	4	1.930	4	4.
Threonine	7	6.965	7	7.
Alanine	6	5.315	6	6.
Tyrosine	1	2.979	1	1.
Histidine	3	1.443	3	1.373
Glutamine	2	3.586	2	3.626
Asparagine	5	5.208	5	6.316
Lysine	18	12.935	18	16.683
Aspartic Acid	3	3.974	3	3.021
Glutamic Acid	8	9.871	8	7.978
Cysteine	2	1.532	2	3.126
Tyrosine	5	2.235	5	3.873
Arginine	2	9.238	2	2.
Glycine	13	7.119	13	13
Chi-Square		31.418		4.148

Acknowledgment

The authors would like to thank Dr. S.E. Fienberg for a number of important comments and suggestions.

References

Chen, T. 1972. Mixed-up frequencies and missing data in contingency tables. Unpublished Ph.D. dissertation, Univ. of Chicago.

Fienberg, S.E. 1970. The analysis of multidimensional contingency tables. Ecology 51, 419-433.

Fienberg, S.E. 1972. The analysis of incomplete multi-way contingency tables. Biometrics 28, 177-202.

Goodman, L.A. 1968. The analysis of cross-classified data. Independence, quasi-independence, and interaction in contingency tables with or without missing entries. J. Amer. Statist. Assoc. 63, 1091-1131.

Kimura, M. and Ohta, T. 1972. Population genetics, molecular biometry, and evolution. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability V, 43-65.

King, J.L. 1972. The role of mutation in evolution. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability V, 69-100.

King, J.L. and Jukes, T.H. 1969. Non-Darwinian evolution: random fixation of selectively neutral mutants. Science 164, 788-789.

Ohta, T. and Kimura, M. 1970. Statistical analysis of the base composition of genes using data on the amino acid composition of proteins. Genetics 64, 387-395.

Ohta, T. and Kimura, M. 1971. Amino acid composition of proteins as a product of molecular evolution. Science 174, 150-153.

Richmond, R.C. 1970. Non-Darwinian evolution: a critique. Nature 225, 1025-1028.